Evaluation of ASR Systems using Information Retrieval Measures

Artem Peregoudov
IDIAP Research Institute, Martigny, Switzerland
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

ABSTRACT

Performance evaluation measures play an essential role in the design of automatic speech recognition systems as they are used to predict the performance of the system in a real application, to compare systems and to analyse the errors introduced during the recognition process. Commonly used word error rate (WER) has proven to be a good measure for most 'old-school' applications, such as dictation systems, where the whole content is of interest and every word has the same importance. However, with advances made in the recent past, more systems are currently emerging, aiming to resolve a new range of tasks. Multimodal systems, spoken document retrieval and call routing are examples of applications where the task involves categorisation and indexing of the audio content and where not all of the words have the same importance anymore. In this work we investigate on the evaluation of several ASR systems on the Ressource Management task using information retrieval measures.

DEFINITIONS / COMPARISON

Word Error Rate

Information Retrieval Measures [1]

both measures are based on an alignment between the reference transcription and therecogniser output

$$WER = \frac{S+D+I}{N_r}$$

N_r: number of words in the reference transcription
S: Substitution count
D: Deletion count
I: Insertion count

Well suited for an entire range of applications such as disctation systems.

Limitations :

- No upper bound
- Suffers from the error overestimation due to the dynamic programming alogrithm used for alignment [2].

Indexed alignment:

Reference:	The	cat	ε	sat	on	the	mat	at	the	door
Recognised:	She	rat	sat	sat	ε	the	mat	at	ε	door
Slot index *j*:	1	2	3	4	5	6	7	8	9	10

- *V* : set of unique words
- ε : null word
- $\forall v_i \in V$:
  - $R_i = \{j \mid r_j = v_i\}$, relevant units in the reference transcription
  - $A_i = \{j \mid a_j = v_i\}$, retrieved information units
  - $R_i \cap A_\iota = \{j \mid r_j = a_i = v_i\}$, correctly retrieved units
- These are used to define per-word precision (fraction of the retrieved information units which is relevant) and recall (fraction of the relevant units which has been which has been retrieved).
- 2 types of averages are defined over the entire vocabulary. The micro-averages weight every information unit, while the macro-averages weight words.

Advantages :

- Application oriented word weights can be introduced in both averages.
- Word based measure
- Strictly bounded to [0, 1]

Per-word recall and precision :

$$\rho_i = \frac{|R_i \cap A_i|}{|R_i|} \qquad \pi_i = \frac{|R_i \cap A_i|}{|A_i|}$$

Micro-averaged recall and precision :

$$\rho_\mu = \frac{\sum_i |R_i \cap A_i|}{\sum_i |R_i|} \qquad \pi_\mu = \frac{\sum_i |R_i \cap A_i|}{\sum_i |A_i|}$$

Macro-averaged recall and precision :

$$\rho_M = \frac{1}{|V_\mathrm{r}|}\sum_i \rho_i \qquad \pi_M = \frac{1}{|V_a|}\sum_i \pi_i$$

Word-weighted averages :

$$\rho_{Mw} = \frac{1}{\sum_{i \supset R} w_i}\sum_i w_i \rho_i \qquad \pi_{Mw} = \frac{1}{\sum_{i \supset A} w_i}\sum_i w_i \pi_i$$

$$\rho_{\mu w} = \frac{\sum_i w_i |R_i \cap A_i|}{\sum_i w_i |R_i|} \qquad \pi_{\mu w} = \frac{\sum_i w_i |R_i \cap A_i|}{\sum_i w_i |A_i|}$$

References
[1] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation", IDIAP-RR 73, IDIAP, Martigny, Switzerland, 2004.
[2] M.J. Hunt, "figures of merit for assessing connected-word recognisers", *Speech Communication*, vol. 9, no. 4, pp. 329-336, Aug. 1990.

EXPERIMENTS

Preliminary results obtained by evaluating 4 ASR systems on the DARPA Ressource Management task (continuous speech). 2 types of systems: HMM/GMM and TANDEM, based on 2 types of subword units : phonemes and graphemes. Currently, ongoing work on the evaluation of a call routing application.

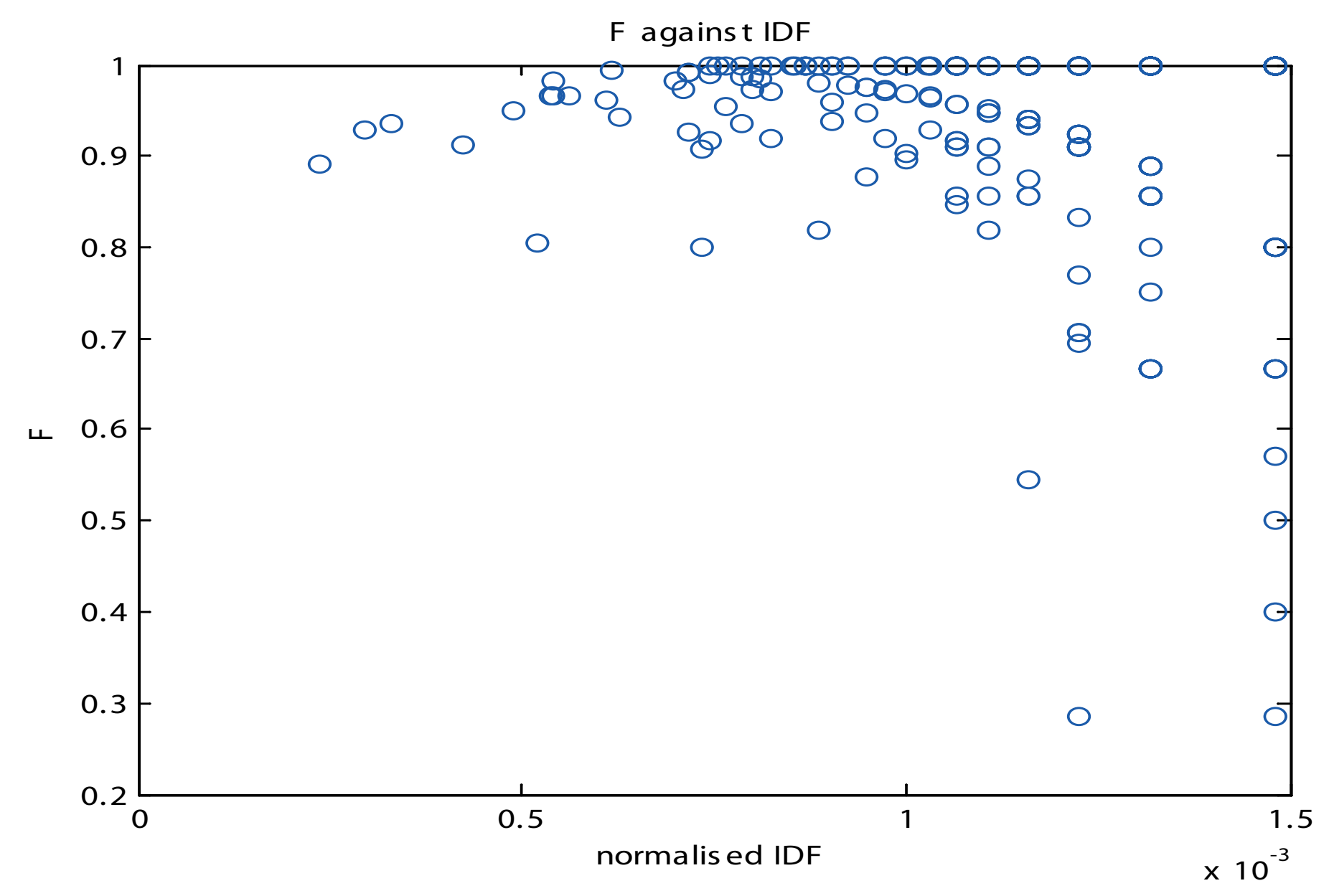
IDF WEIGHTING

IDF (Inverse document frequency) is a measure showing how a word is discriminant with respect to a document set (in our case the documents are the different utterances)

For a word appearing in n documents out of a total of N :

idf = log2(N/n)

weighting scheme	all 1	idf	all 1	idf	all 1	idf	all 1	idf
	GMMGraphemes		GMMPhonemes		TANDEMGraphemes		TANDEMPhonemes	
micro-F	92.2%	92.5%	94.0%	94.3%	94.1%	94.6%	94.7%	95.0%
MACRO-F	91.9%	92.8%	93.4%	94.1%	94.2%	95.0%	94.6%	95.1%
WRR (1-WER)		90.1%		92.4%		92.6%		93.2%

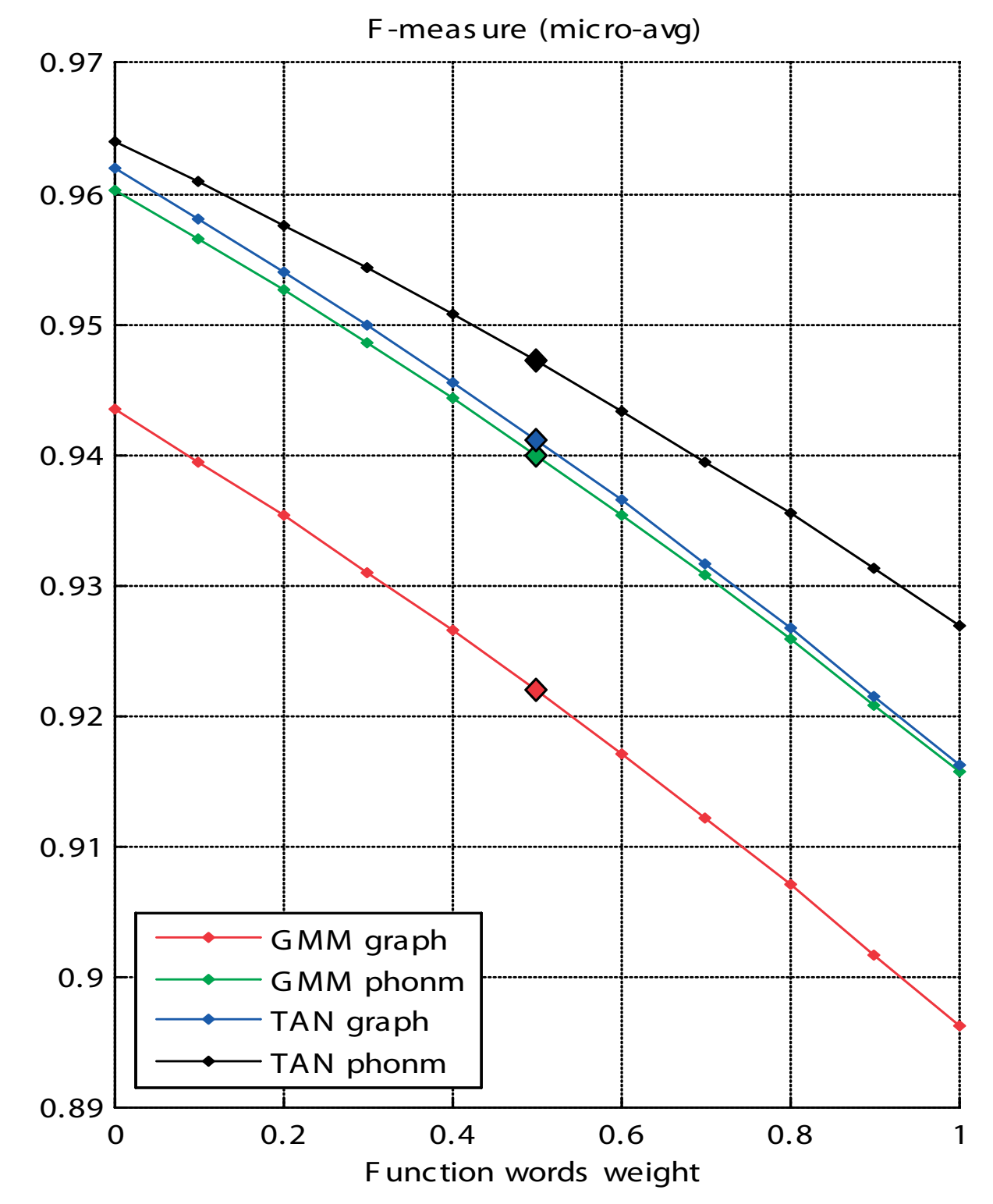


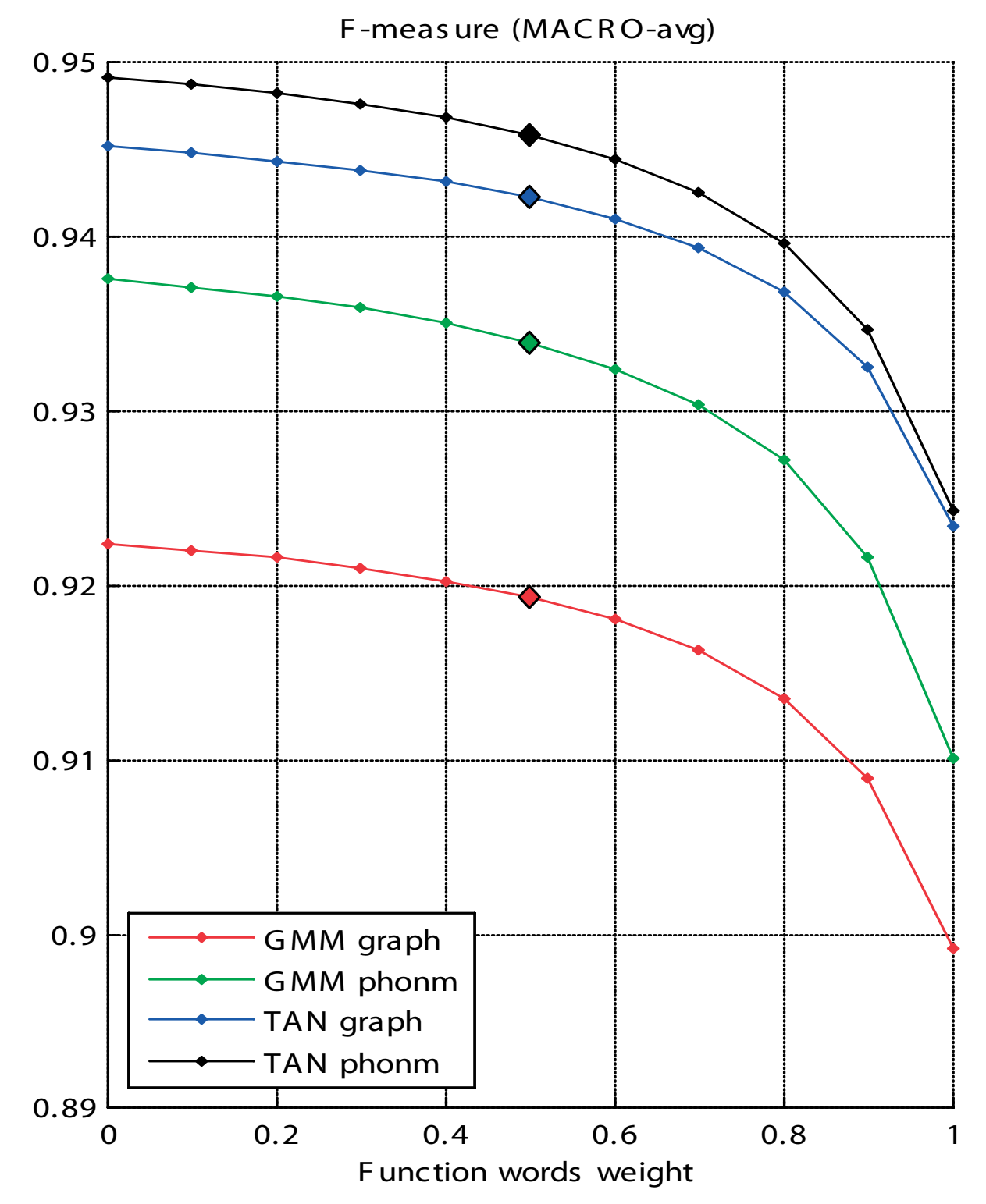
F-measure against IDF for TANDEM Phoneme system

FUNCTION WORDS WEIGHTING

Function words obtained from the stop list used by the Idiap Text Retrieval system:

- 390 out of the 990 unique words of the vocabulary
- 4656 out of the 10288 words appearing in the utterances
- Function weights w_f varying from 0 to 1, non-function weights w_{nf}=*1*-w_f





CONCLUSIONS / ONGOING & FUTURE WORK

- IR measures still suffers from the error underestimation due to the dynamic programming alignment. Looking for alternatives.
- Evaluation of call routing application
- Evaluation of a multimodal ASR system

ACKNOLEDGEMENTS
Big thanks to Mathew Magimai Doss (IDIAP) for his results on the RM task and his wise advices, to Jean-Yves Parfait (IDIAP) for his results on the call-routing application, to John Dines (IDIAP) and Alessandro Vinciarelli (IDIAP) for the fruitful discussions we shared.